DIG64 Technical Whitepaper: IA-64 Server Storage Solutions

Intel Corporation
October 1999

Mark Bradley, Adaptec
Duane Grigsby, QLogic
Jin-Lon Hon, Mylex
Julie Oehler Schott, LSI Logic

Overview

This whitepaper discusses some of the available mass storage technologies for high-end server systems, and recommends which technologies should be incorporated into IA-64 based servers. This whitepaper is one of a series of technical whitepapers supporting the *Developer's Interface Guide for IA-64 Servers* (DIG64).

This document presents information developed by the DIG64 storage subgroup. However, it does not specify DIG64 compliance requirements. Refer to the *Developer's Interface Guide for IA-64 Servers* for the complete compliance requirements. The information within this paper assists in the selection and configuration of storage solutions for IA-64 servers, to ensure their ease of use and to help provide reliable, fast storage solutions for IA-64 server customers.

IA-64 systems used as servers must have reliable, high-performance storage subsystems to provide the overall system characteristics that enterprise computing and data center users require. High-performance storage is critical to the proper functioning of servers, because it is the data repository for users and database applications, and it is typically the system's boot image source. Slow, unreliable data storage subsystems lead to servers with poor performance and reliability characteristics.

At present, SCSI and Fibre Channel FCP are recognized as the high-performance storage interfaces for IA-64 servers. These interfaces provide high data transfer rates, connectivity for many peripherals, proven technology, flexible configurations, and wide availability that will allow them to be used for the foreseeable future.

ATAPI/IDE are not discussed in this whitepaper. This interface is not considered appropriate for primary storage in IA-64 based server systems for performance and connectivity reasons.

Manufacturers looking to select high-performance components for server systems should look for the following design features in storage components:

- Adapter and controller devices that are compliant with PCI 2.1 or later.
- SCSI or Fibre Channel controllers are used as the primary storage interface.
- RAID using SCSI and Fibre Channel storage interfaces.
- Primary storage adapters that are usable by the host computer system to boot the host computer from disk.
- PCI or PCI-X adapters and controllers with bus mastering and burst mode DMA capabilities.
- High-performance peripheral devices (i.e., fast data access and high data transfer rates for the storage devices).
- Adapters and controllers that have 64-bit enabled device drivers for the operating system.
- Adapters that provide scatter/gather capability.
- High-quality power, cabling, and cooling for storage components.

Small Computer System Interface

Small Computer Systems Interface (SCSI) has been in use as a standard in the industry since 1986. A SCSI adapter provides an interface for mass storage devices, such as magnetic disk devices, tape drives, optical disk drives, and CD class devices. SCSI also supports media changers for many of the removable media devices. SCSI peripherals are discussed later in this whitepaper.

For this discussion, a host adapter is either a host bus adapter or *controller* that is inserted into the system's expansion bus, or it is a discrete component designed directly into a system baseboard local bus. Either configuration allows the SCSI logic to communicate, via the system bus, with the host memory.

The first industry standard for SCSI was SCSI-1. This interface allows up to eight devices (seven plus one controller) to be connected over a parallel, 8-bit data

path. Each of the seven addresses supports up to eight logical unit numbers (LUNs), which are subunits of the target device. An example of a target device having a LUN is a multidisk CD-ROM changer. For example, the CD-ROM device may have a target ID of 3, and the first disk of the CD-ROM device has a LUN of 0, and second disk has a LUN of 1, and so on. With SCSI-1, signal lines are single-ended and open collectors are used to drive the bus. Command and data transfers are asynchronous and the maximum data transfer rate is 5 MB per second.

SCSI-2 incorporates many improvements over SCSI-1. SCSI-2 (often referred to as fast SCSI) doubles the frequency of the bus from 5 MHz to 10 MHz, thereby doubling the maximum transfer rate to 10 MB per second. Note that SCSI does not incorporate a specific bus clock signal; the bus frequency is determined by the minimum cycle time.

SCSI-2 also allows expanding the data path from 8-bits to 16-bits or 32-bits, thereby increasing the maximum transfer rate again to 20 MB or 40 MB per second. Expansion of the bus width also allows addressing of up to 16 devices on a single bus. Other improvements include the ability to send data synchronously (commands are asynchronous), rather than requiring an acknowledgment for each data cycle. SCSI-2 also permits single-ended or differential signaling (called high-voltage differential—HVD in SCSI-2). Differential signaling allows for greater noise immunity and SCSI cable distance and improved transfer rate. Lastly, SCSI-2 described what is called the Common Command Set, which comprised a basic set of commands designed to improve interoperability between adapter and peripheral manufacturers.

While SCSI-1 described only the hardware interface bus and SCSI commands in general, and SCSI-2 specifications captured the hardware interface and the Common Command Set, SCSI-3 divides the specification into several stand-alone documents, with each document addressing a particular layer of the interface. These documents cover the SCSI-3 Parallel Interface (SPI), the SCSI Interlock Protocol (SIP), the SCSI-3 Architectural Model (SAM), and the SCSI-3 Primary Commands (SPC). Additional documents describe the specific protocols and command sets to support specific types of peripherals and functionality.

In addition to numerous protocol and peripheral support enhancements defined in the named specifications, SPI increases the bus frequency to 40 MHz and also implements “double transition clocking” where data is clocked into the receiver at both the leading and trailing edge of the REQ or ACK signal. SPI also supports low-voltage differential (LVD) signaling, opposed to HVD in SCSI-2. HVD has power requirements that often prevented the signal drivers from being implemented on the controller chip. In contrast, LVD allows these drivers to more easily be placed in the controller chip, thereby reducing cost. The maximum data transfer rate for SCSI-3 SPI is 160 MB per second.

In addition to the SPI, SCSI-3 also supports three new physical interfaces. They are Fibre Channel, Arbitrated Loop (FC-AL), Serial Storage Architecture (SSA), and P1394. FC-AL supports a maximum data transfer rate of 100 MB per second, SSA supports a duplex interconnect system providing 20 MB per second simultaneously in both directions, and P1394 supports a data transfer rate of 100 MB per second and higher.

SCSI Adapters

The following paragraphs describe some of the SCSI-based features and functionality that are necessary in IA-64 systems.

PCI-SCSI adapters should support PCI bus mastering. This capability allows the SCSI adapter residing on the PCI bus to perform DMA operations into DRAM memory, thereby freeing the system CPU(s) from the necessity of performing these operations directly. PCI bus mastering should be enabled by default on all SCSI adapters.

DMA and scatter/gather capability should also be enabled for all SCSI adapters. Burst mode DMA should be enabled by default on SCSI adapters for performance reasons. The only exception is during the boot phase. From boot to runtime, operation mode switching should be invisible to the user and the host OS. This

means that the adapter or its driver should perform the mode switching without external intervention.

SCSI bus termination is required to reduce signal reflections caused by an impedance mismatch existing between the characteristic impedance of the transmission line and the endpoints. Termination is applied at the extreme endpoints of the bus. Typically, SCSI bus termination is applied manually by the person configuring the SCSI system, and the process is error prone, which results in poor system reliability. Conversely, automatic termination is a feature where the SCSI controller senses the need for termination and automatically enables it if required. This allows a user to add external SCSI devices without removing the server covers and manually changing termination.

In IA-64 servers, SCSI controllers should provide automatic termination. Terminators used in the SCSI host adapter should be regulated terminators, which are also known as active, SCSI-3 SPI, SCSI-2 alternative-2, or Boulay terminators. SCSI termination built onto internal cables should meet approved SCSI specifications defined in *Small Computer System Interface-3 (SCSI-3)* ANSI X3-253. All terminators on the external SCSI bus should be powered from the TERMPWR lines in the SCSI bus.

The circuit supplying TERMPWR should have built-in overcurrent protection. Devices that provide TERMPWR should also provide a way to limit the current through use of a self-resetting device. (For example, a positive-temperature coefficient device or circuit breaker can be designed into the circuit. These devices open during an overcurrent condition and close after the condition ends.)

The SCSI-3 specification allows for both differential and single-ended signal transmissions over a single SCSI bus although only one transmission mode is permitted to operate on any given continuous bus. As described previously, SCSI-3 SPI specifies Low Voltage Differential (LVD) transmission; it does not address High Voltage Differential (HVD). The SCSI controller must sense LVD differential signaling (DIFFSENS) capabilities on the bus and determine if any agents on the bus are capable of only single-ended transmissions. If so, the appropriate action must be taken to avoid damaging drivers/receivers on the connected devices.

All SCSI peripherals and the SCSI host adapter should implement the SCSI bus data protection signals (parity) defined in the SCSI-3 specification. Data protection should be enabled by default.

Some SCSI systems are used in fault-tolerant server systems or other configurations, where multiple processors access shared peripherals. A feature provided by SCSI, called multi-initiator support, allows two or more SCSI adapters or controllers to coexist on a shared SCSI bus with peripheral devices. If the SCSI controller is to be used on a shared SCSI bus, then each SCSI controller has to provide multi-initiator support. (Here, a shared SCSI bus is a SCSI bus with more than one SCSI adapter/controller and a number of peripherals other than zero). This facilitates possible device sharing using SCSI *reserve* and *release* mechanisms, as described in SCSI-3 specifications.

For use in a system intended as a node in a cluster using shared SCSI, the SCSI IDs should be changeable from the default SCSI controller. Also, the boot time SCSI bus reset operation should be disabled on each controller attached to a shared bus.

A wide array of cables, connectors, and terminations are used in servers, leading to confusion as to what exactly goes where. This is especially true for a technician performing a system upgrade. Electrical and physical system damage can result. Connectors, SCSI adapters, peripherals, cables, and terminators should be clearly labeled to show the bus type. All external SCSI connectors should display the appropriate SCSI icon defined by X3T10 in addition to the electrical signaling acronyms listed below.

- **DIFF**—differential. A signal type used in external large storage cabinets, also called HVD.
- **SE**—single-ended. The most commonly used signal type, such as those found in home PCs and high-end workstations.

- **LVD**—low-voltage differential. A signaling method similar to DIFF but with lower signaling voltages supporting higher transfer rates.

In IA-64 server systems, 50-carrier internal connectors for SCSI should be ribbon-style connectors. Sixty-eight-pin internal connectors should be sub-miniature 'D' and pinouts should conform to the specification *Small Computer System Interface-3 (SCSI-3)* ANSI X3-253. External connectors should provide metallic shielding for RF and conducted emissions reasons.

For internal and external configurations, the SCSI bus cable should be plugged into shrouded and keyed connectors on the host adapter and devices to avoid incorrect plugging. For internal configurations, pin 1 orientation should be designated on one edge of the ribbon cable and also on the keyed connector for the SCSI peripheral device.

To avoid possible confusion, other devices in the server should not use SCSI-type external or internal connector types, even if they are distinctly labeled or marked.

All adapter products should be certified to FCC-B, VDE/TUV, and Asian compliance agencies (e.g., VCCI). Compliance reports should be available at the time of use of an adapter in an IA-64 computer system. The marking of computer components of this type are covered by the regulating agencies and these markings are required on SCSI adapters, as per these laws. Form factor for various bus adapters is implementation-specific, depending on the particular I/O bus used in the server. The PCI SCSI adapter form factor should adhere to PCI Local Bus Specification, Revision 2.2. Power consumption for adapters should meet PCI Local Bus Specification, Revision 2.1, Revision 2.2, and PCI Bus Power Management Interface Specification.

SCSI Peripherals

All SCSI peripherals should be UL certified and should be tested and appropriately marked as compliant with FCC-B, VDE/TUV-B, and Asian compliance certified for EMI/RFI to level B compliance.

A valuable feature, in terms of heat generation and cooling, power consumption, and unit life, is the ability to place equipment and devices in a "low-power state" when the equipment is not in use. For disk drives, the SCSI-3 specification supports the ability to spin down the drive and thereby consume less power. This feature is called START/STOP Unit Support. For IA-64 systems, this feature is required, and it is applicable to SCSI hard drives, CDs, and other optical devices. When the SCSI device supports the START/STOP commands, the behavior of the commands should be as defined in *ANSI NCITS T10 SCSI-3 Multi-Media Command Set-2 (MMC-2)*. Operating system policy and the policy of certain management agents determine when and if to cause a drive to stop and/or start.

Hot Plugging allows for the removal and replacement of SCSI devices while the server is still powered up and running. SCSI adapters and peripherals in IA-64 servers should meet the removal and insertion requirements of SCSI devices as defined in Case 4 of the SCSI-3 specification. A locking mechanism is recommended to ensure that devices are not removed when they are in use.

SCSI magnetic hard disk drives should conform to the specification *Small Computer System Interface-3 (SCSI-3)* ANSI X3-253 and should transfer data on the SCSI bus at a minimum rate of 10MB/sec. The following hard disk drive features are recommended:

- Tagged command queuing
- Read ahead enabled
- Write immediate capability
- Rotational speed of 7200 RPM or greater
- Average seek of 15 msec. or less, minimum seek time of 3 msec. or less, maximum seek time of 25 msec. or less
- 3.5-inch, half height form factor for internal drives, external drive not specified
- MTBF of 250,000 hours or more

- Power consumption of 20 watts or less during typical operation
- Operating range of: 5-50o C; 5-90% RH, noncondensing; 10G shock, 0.5G vibration
- Predictive failure mechanism

Tape drives are the devices typically used for archival backup of hard drives and hard drive-based subsystems. At least one SCSI tape drive per system should be provided for minimal backup and data exchange. SCSI tape drives that will satisfy this requirement are listed below in preferential order, with capacity, performance, and media interchange as the ranking criteria.

- digital linear tape, 30 GB native capacity or more
- digital audio tape (DDS-3 or -4 compliant), 12 GB native capacity or greater
- 8mm helical scan tape, 7 GB native capacity or more

Generally, it is recommended that the data compression features these drives provide be enabled, as compression algorithms are generally reliable and interchange problems due to compression algorithm differences generally have been resolved.

Tape drives, because of the flexible and varied media, have diverse performance, operating, and power specifications. These will remain unspecified at this time.

Optical disk drives, including WORM (write once, read many) and MO (magneto optical), often are used for intermediate storage (near-line storage). While performance typically is not as good as hard drives, it may be acceptable for files that are not considered performance critical for system applications. Characteristics of optical drives are low error rates ($<10^{-13}$ BER), random access with average seek time on the order of 18–35 milliseconds and 2–3 MB/sec. transfer rates. If incorporated by the server, WORM media should meet ANSI X3.200-1992 (R1997) Information Systems - Unrecorded Optical Media Unit for Digital Information Interchange and ISO/TR 12037:1998. Worm drives should support these standards for optical media. Optical drives do not have standard form factors, or consistent general performance indices, environmental, or power specifications. These will remain unspecified at this time.

A CD-ROM that reads ISO-9660 may be needed in IA-64 servers for software distribution. If this CD-ROM is not accessible over the server's network, then there should be a CD-ROM in each server. The minimum drive speed should be 8X. The recommended characteristics of this CD-ROM driver are as follows:

- Caddy-less models only can be used
- 5.25-inch half high form factor or less (industry standard form factors only are acceptable)
- CD- CD-Erasable (rewritable)

Media changers exist for many types of removable media devices. At this time, this includes CDs, optical disks, and tapes. Both *Independent Medium Changers* and *Attached Medium Changers* are acceptable for use on IA-64 systems. Media changers are not required for IA-64 systems, but if such changers are used, they should comply with the SCSI-3 MCC (Media Changer Commands) profile of ANSI X3T10. These include exchange medium, initialize element status, move medium, position to element, read element status, release element, request volume element address, and reserve element.

Fibre Channel

Fibre Channel is a high-speed transfer interface technology that maps several common transport protocols, including IP and SCSI, allowing it to merge high-speed I/O and networking functionality in a single connectivity technology. Fibre Channel is an open standard as defined by ANSI and OSI standards and operates over copper and fiber-optic cabling at distances of up to 10 kilometers. It is unique in its support of multiple interoperable topologies, including point-to-point, arbitrated loop, and switching, and it offers several qualities of service for network optimization. With its large packet sizes, Fibre

Channel is ideal for storage, video, graphic, and mass data transfer applications.

The command protocols include SCSI-3, IP, HIPPI, and others. Fibre Channel interface enables both I/O channel protocols such as SCSI-3 and network protocols such as IP on the same serial interface.

A Fibre Channel host adapter is a host bus adapter or *controller* that is inserted into the system's expansion bus, such as a PCI bus slot. For the purposes of this whitepaper, it may also be a discrete component designed directly into a system baseboard PCI bus or other such bus. This allows the Fibre Channel transport to communicate to the computer's CPU and memory.

The Fibre Channel specifications define standards and guidelines, in terms of profiles, for the physical interface, the topologies, and the command set protocols. The physical interface defines the physical characteristics of the media, including cables, connectors, and component dimensions. The topology standard defines different schemes to interconnect one or more devices; examples include point-to-point, arbitrated loop, and switched fabrics. The command set protocols define different protocols an application could use to transport its data.

Fibre Channel connects to various peripherals, including disk drives, tape units, tape libraries, storage subsystems, graphics terminals, and high-speed laser printers. Fibre Channel adapters and peripherals provide high-performance servers with a method of connecting both network and storage devices, but this whitepaper addresses only storage.

Fiber Channel technology provides high-performance and large connectivity at long distances. Depending on the physical medium and other issues, Fibre Channel can support data transfer rates up to 1.0625 GB/second. Currently, interconnections of up to 126 devices or nodes are supported.

The Fibre Channel standard defines a five-layer protocol architecture. The higher layers define mappings from other communications protocols onto the Fibre Channel fabric. Supported protocols include:

- Small Computer Systems Interface (SCSI)
- Internet Protocol (IP)
- ATM adaptation layer for computer data
- Link Encapsulation
- IEEE 802.3
- Virtual Interface Architecture (VIA)

All the supported protocols listed above can be used simultaneously. For example, a Fibre Channel arbitrated loop running IP and SCSI protocols can be used for both system-to-system and system-to-peripheral communication, eliminating the need for separate I/O controllers.

The data transmission mechanism, the different types of physical media, and the data rates supported by Fibre Channel are defined in the following specifications:

- Fibre Channel Physical and Signaling Interface (FC-PH) X3.230-1994
- Fibre Channel Physical and Signaling (FC-PH-2) X3.297
- Fibre Channel Physical and Signaling (FC-PH-3) X3.303

These combined specifications define the lower levels of the Fibre Channel transport. Fibre Channel adapters should be compliant with these standards. However, certain profiles discussed herein may eliminate or restrict some feature(s) of the standard.

Fibre Channel defines a standard topology for devices that are shared on a single physical loop. The Fibre Channel adapter should be compliant with the Arbitrated Loop (FC-AL) X3.272-1996 and should also be compliant with working draft Arbitrated Loop (FC-AL-2) project 1133-D. Fibre Channel as a host interface should support Fibre Arbitrated Loop (FC-AL) for shared arbitrated loops.

Fibre Channel defines a standard topology for connecting to devices through a switch or point-to-point.

Adapters that support switched topologies should be compliant with Fabric Generic Requirements (FC-FG) X3.289 and Switched Fabric (FC-SW) T11/0959-D.

Fibre Channel as a host interface should support Fibre Channel Switch Fabric (FC-SW) for public loop and point-to-point protocols. As mentioned previously, the Fibre Channel protocol supports FCP-SCSI for transporting the SCSI command set over Fibre Channel transport. Fibre Channel adapters that support the SCSI command set should be compliant with the Fibre Channel protocol (FCP) X3.269: 1996, and could be compliant with working draft Fibre Channel protocol (FCP-2) project 1144-D. Fibre Channel as a host interface should support Fibre Channel IP protocol (FC-IP) for adapters, drive enclosure, or RAID subsystems.

Fibre Channel defines a subset of the standard that describes features supported by direct-attach disk drives on the shared Fibre Channel loop. Adapters that support this profile should be compliant with the working draft of Private Loop Direct Attach profile (FC-PLDA) T11/project 1162-DT.

Fibre Channel defines a subset of the standard that details features supported by tape units on the Fibre Channel loop. Adapters that support attach tapes should be compliant with working draft FC-TAPE profile T11/project 1315-DT.

Fibre Channel defines a standard subset of features that adapters/devices should support for switched fabrics attached on a loop. Adapters that support switched fabrics should be compliant with the Fabric Loop Attachment profile (FLA) T11/1235-DT.

PCI-Fibre Channel Adapters

It should be noted that only PCI-Fibre Channel adapters are considered in this whitepaper. A host adapter residing on the PCI bus must be compliant with PCI Local Bus Specification, Revision 2.2 or later. External controllers with their own enclosures and peripherals are not covered herein, as these are typically proprietary products.

PCI-Fibre Channel adapters should support PCI bus mastering. This capability allows the Fibre Channel adapter residing on the PCI bus to perform DMA operations into host DRAM memory, thereby freeing the system CPU(s) from the necessity of performing these operations directly. PCI bus mastering should be enabled by default on all Fibre Channel adapters.

DMA and scatter/gather capability should be enabled on all PCI-Fibre Channel adapters. Burst mode DMA should be enabled by default on Fibre Channel adapters for performance reasons. PCI-Fibre Channel adapters should support the PCI bus power management interface specification. This includes correct implementation of the PCI configuration space registers used by power management operations, and the appropriate device state (Dx) definitions.

PCI-Fibre Channel adapters should be compliant with the standards specified in FC-PH for cables and connectors. Connectors with Fibre Lasers should be rated Class 1.

- Fibre optic cables - SC Duplex Multimode
- Fibre arbitrated loop cables - DB9 and HSSDC connectors

It is required that form factor and power consumption for PCI-Fibre Channel adapters adhere to *PCI Bus Power Management Interface Specification, Revision 1.1.* Connectors, Fibre Channel adapters, peripherals, cables, and fibre should be clearly labeled to show the bus type. Fibre laser should be clearly labeled with **TX** and **RX**.

Fibre Channel Peripherals

All Fibre Channel peripherals should be UL certified and should be tested and appropriately marked as compliant with FCC-B, VDE/TUV-B, and Asian compliance certified for EMI/RFI to level B compliance.

A valuable feature, in terms of heat generation and cooling, power consumption, and unit life is the ability to place equipment and devices in a “low-power state”

when that equipment is not in use. For disk drives, the SCSI-3 specification supports the ability to spin down the drive and thereby consume less power. This feature is called START/STOP Unit Support. For IA-64 systems, Fibre Channel hard drives should conform to *ANSI NCITS T10 SCSI-3 Multi-Media Command Set-2 (MMC-2)* or later versions of the ANSI X3T10 specification. This feature is applicable to SCSI hard drives, CDs, and other optical devices. Operating system policy and the policy of certain management agents determine when and if to cause a drive to stop and/or start.

Hot Plugging allows for the removal and replacement of a device while the server is still powered up and running. Fibre Channel devices in IA-64 systems should meet the removal and insertion requirements of such devices as defined in Case 4 of the -3 specification. A locking mechanism is recommended to ensure that devices are not removed during critical times of use.

Dual port with regard to Fibre Channel means that the peripheral subsystem is connected via two physical Fibre Channel connections to two distinct loops, providing a secondary access method to the peripheral subsystem in event of failure. Fibre Channel peripherals should provide dual-port support for nodes sharing the same device in such clustering applications.

The following features for Fibre Channel hard-disk drives are recommended:

- Tagged command queuing
- Read ahead enabled
- Write immediate capability (desirable for some applications; not recommended for others unless data is protected from power loss or hardware failure)
- Rotational speed of 7200 RPM or more
- Average seek of 15 msec. or less; minimum seek time of 3 msec. or less; maximum seek time of 25 msec. or less
- 3.5-inch half height form factor for internal drives; external drive not specified
- MTBF of 250,000 hours or more
- Power consumption of 20 watts or less during typical operation
- Operating range of: 5-50oC; 5-90% RH, noncondensing; 10G shock, 0.5G vibration
- Predictive failure mechanism (i.e., S.M.A.R.T.)

For IA-64 system, Fibre Channel tape subsystems are optional devices that may be used for archival backup of hard drives and hard drive-based subsystems. Fibre Channel tape specification should be complied with for such devices.

Redundant Array of Independent Disks

RAID stands for Redundant Array of Independent Disks. According to the RAID Advisory Board (RAB), RAID is a disk array in which part of the physical storage capacity is used to store redundant information about user data stored on the remainder of the storage capacity. The redundant information enables regeneration of user data in the event that one of the array's member disks or the access path to the disk fails.

RAID uses data parity, data stripping, data mirroring, and error detection and correction algorithms to provide for overall improved throughput and fault tolerance.

Data stripping is the process of splitting the data across two or more disks at the bit, byte, or block level. For example, in a four-drive system, the first block is written to drive one, the second block is written to drive two, and the third block to drive three, and fourth block to drive four. The fifth block will then be written to the second block location at drive one. The sixth block is written to the second block location at drive two and so on.

Data mirroring repeats the same data image in a different drive. Data parity technique generates a parity data corresponding to user data on several drives in the form of bit-by-bit parity. The parity data

RAID Level 0	This level uses more than one drive and simply performs data stripping at the bit, byte, or block level. Access to the individual drives occurs in parallel. The purpose is to improve the overall throughput of the storage system because data access occurs in parallel simultaneously across multiple drives. However, without data overlap or redundancy, this level offers no protection against a single drive failure.
RAID Level 1	This level uses two equal capacity drives that mirror each other. One disk duplicates all the files of the other and serves as a backup disk. No data stripping occurs.
RAID Level 2	This level incorporates data stripping as defined for Level 1 but also uses redundant disks to correct single-bit errors and detects double-bit errors.
RAID Level 3	This level uses multiple drives to perform data stripping but uses only parity for error checking. This level typically requires only one extra drive for error detection, versus perhaps three or more for error detection and correction in Level 2.
RAID Level 4	This level performs data stripping at the sector level. Sectors are read serially from the first drive and the second drive, and so on. An additional drive is used for parity checking.
RAID Level 5	This level is similar to Level 4, except that the parity information is written to the data drives, and the dedicated parity drive is removed.
RAID Level 6	This level is the same as Level 5, except that an additional parity drive is added. This level is more tolerant of drive failure.

together with user data is stripped over several drives. With data parity or data mirroring, any single failed drive or drive interface still allows for the data to be recovered from surviving drives.

The RAB currently recognizes seven RAID implementation levels in its RAID book, 6th edition. The following table is an overview of the seven levels currently described by the RAB.

Some manufacturers may have different definitions for these levels. Some vendor-unique RAID levels or a combination of multiple RAID levels, such as RAID 0 and 1 (commonly referred to as RAID level 10), are widely used in the industry.

Although RAID Level 0 does not have data-redundant protection, it is in common use and the RAB endorses the term RAID Level 0 to refer to disk stripping.

For IA-64 systems, RAID can be applied to SCSI, Fibre Channel, or other disk and tape technology. The host adapter attachment may be PCI bus, SCSI bus, Fibre Channel, or a combination of these. The peripheral device interface may be SCSI or Fibre Channel. Both RAID types may have SCSI or Fibre Channel as the back-end device interface. Peripheral information presented in the SCSI Peripherals and Fibre Channel Peripherals sections of this document is applicable to these controllers as well. Attachment of devices to RAID controllers does not imply their exclusive use as RAID devices.

Host-Based RAID Controllers

A host-based RAID controller is a controller card that is inserted into the system's expansion bus, such as a PCI bus slot. For the purposes of this whitepaper, it may also be a discrete component or set of components designed directly into a system baseboard PCI bus or other such bus. Such a configuration allows the SCSI or Fibre Channel logic to use that bus to communicate

with the computer's CPU or memory. The RAID controller must be compliant with *PCI Local Bus Specification 2.1* or later with Subsystem ID and Subsystem Vendor ID support.

For IA-64 systems, the RAID controller should support PCI bus mastering. Bus master capabilities should meet the related specification for the particular controller. The RAID controller also should use PCI advanced commands. These are *memory read multiple, read line, write multiple,* and *memory write & invalidate* for optimum PCI bus efficiency. The RAID controller also should support Dual Address Cycle (DAC) for greater than 4GB memory access.

If PCI power management is implemented, it should be compliant with the *PCI Bus Power Management Interface Specification, Revision 1.1* or later. To support hot swap functionality, the RAID controller should support the PCI Hot-Plug specification 1.0 for replacing field replaceable units (FRU) without interrupting the system availability. The card should conform to the PCI Hot-Plug requirements and provide device driver or software support per the resident operating system.

The RAID controller should support RAID Level 0 (data stripping), RAID Level 1 (data mirroring), RAID Level 5 (data with distributed parity), and single disk (JBOD). RAID Level 3 (parallel transfer disks with dedicated parity drive) support for performance improvement for large, sequential IO requests is optional. Support for RAID Level 6 (multiple additional redundant disk), multiple RAID level combinations such as RAID Level 0 plus RAID Level 1, RAID Level 1 plus RAID Level 5, and automatic RAID level adjustment according to host requests is optional.

A RAID controller should have the capability to support disk enclosure management by periodically monitoring disk enclosure temperature, fan, power supply, drive status, and sending appropriate normal or fault status to enclosure indicators such as LED or display panels. The controller should support the SAF-TE specification for a RAID controller with SCSI as the back-end device channel or SCSI Enclosure Service (SES) specification for RAID controller with Fibre Channel as the back-end device channel. Notification of a failed drive is required. The RAID controller should monitor temperature and power voltage, and send notification to the system administrator in the case of a failure event.

In a fault-tolerant RAID configuration, such as RAID Level 5 or Level 1, a failed drive will cause the RAID group to operate in a degraded mode. For IA-64 systems, the controller should regenerate user data from the remaining drives while the system is online. In a degrade mode, one or more drive failures could render the RAID drives into nonoperable mode and cause data loss. Once the failed drive is replaced with a good drive by means of hot-swapping the drive or a hot standby drive, the controller should rebuild the data into the new drive. This is referred to as rebuild mode. Online data rebuild requires that during data rebuild process, the controller continue service to the host system IO request without interruption.

Device hot-swapping support allows the replacement of a failed or failing drive with a good drive while the system is performing its normal function. Once the drive is replaced, a data rebuild process can be automatically or manually initiated.

A standby drive sometimes is referred to as a hot spare drive. Standby drive(s) can be globally assigned to all fault-tolerant RAID groups or assigned to a particular RAID group. When the RAID group has a failing drive, the controller automatically will offline the failed drive and make the standby drive a member of the RAID group. User data will be rebuilt to the new drive automatically without user intervention. Once rebuild is done, the RAID group is brought back to a fault-tolerant state.

RAID controllers that support cache and write-back operation, have to be able to flush cached data and to disable write-back cache.

Drive roaming capability allows a member of a RAID group to change its SCSI ID or physical ID in a Fibre Channel arbitrated loop within channels in the same RAID controller or limited to the same channel. The RAID controller should automatically recognize the RAID group member order in a stripe sequence when

constructing user data. For SCA connector-based drive enclosure, the SCSI ID or loop ID is not jumped in the drive itself and depends on the drive slot inserted. For ease of use, the user should not be required to remember which drive goes to which drive slot when swapping drives.

The controller should support the S.M.A.R.T feature for device failure prediction. This will give early warning for drive replacement and prevent drive failure, thus maintaining performance and fault tolerance levels. Mode 6 is recommended for periodically monitoring drive health condition.

RAID capacity expansion usually is defined as increasing RAID volume capacity by adding additional drives to existing disk arrays while maintaining the same RAID level. Another mechanism is to add new drives combined with a RAID volume addition and/or a RAID-level migration. This capacity expansion can be done online or offline.

RAID-level migration is defined as changing existing RAID level to enhance performance and increase fault tolerance or capacity using the same set of disk arrays while preserving user data. The RAID migration can be done online or offline. IA-64 systems should support RAID level migration.

Removable disk devices are sometimes used as RAID targets. Examples are CD-ROM, DVD, removable media disk, Tape, DLT, Jukebox, and so on. These devices are typically not usable as online storage. In a server platform, it is recommended to restrict the use of these devices to RAID levels 0 and 1. These often are used as a means for software installation (CD-ROM, DVD) and cost-effective local data backup (Tape, DLT).

The RAID controller should support shared device through back-end SCSI or Fibre Channel device connection. The shared device can be accessed from multiple RAID controllers. The shared device capability enables multiple-node clustering support. RAID volume configuration, access right, fault management, and enclosure management are communicated and administrated by all clustered controllers.

RAID Subsystem

In this discussion, a RAID subsystem is a disk subsystem that includes a built-in RAID controller and enclosure(s)—housing multiple drives. This subsystem is different from a JBOD drive box, where no RAID controller is included. For the purpose of this whitepaper, the RAID subsystem interfaces to a host system via SCSI or Fibre Channel. Its back-end device interconnect channel can be SCSI or Fibre Channel. For a RAID subsystem that uses SCSI bus for back-end device interconnects, see the SCSI Adapter section in this whitepaper. For a RAID subsystem that uses Fibre Channel for back-end device interconnects, see the Fibre Channel Adapter section in this whitepaper.

Host System Interface

For a RAID subsystem that uses SCSI bus for host system interconnects, the system interface must conform to SCSI Parallel Interface specification *Small Computer System Interface-3 (SCSI-3)* ANSI X3-253.

For a RAID subsystem that uses Fibre Channel for host system interconnects, supports SCSI for upper-level protocol, and conforms to ANSI SCSI standard for class 2 and class 3 Fibre Channel profiles and services, the system interface must conform to the following:

- *Fibre Channel - Physical Signaling Interface (FC-PH) ANSI X3.230*
- *Fibre Channel Arbitrated Loop (FC-AL-2), Revision 4.25*
- *Fibre Channel Private Loop Direct Attach (FC-PLDA), Revision 2.1*
- *SCSI Fibre Channel Protocol (SCSI-FCP), Revision 12*

Basic RAID functional requirements for a RAID Subsystem are the same as the requirements for a host bus-based RAID controller.

The controller should support multiple initiator in its host interface channel. Multi-initiator support allows the RAID subsystem to be connected to multiple host

adapters—each installed in a separate or the same computer system. This allows multiple data paths to the RAID controller for a resilient and fault-tolerant storage configuration. Multi-initiator support enables multi-nodes clustering support.

Dual port with regard to Fibre Channel means that the RAID subsystem is connected to the host via two physical Fibre Channel connections to two distinct loops, providing for a secondary access method to the peripheral subsystem in case of failure. Fibre Channel RAID subsystem should provide dual-port support for nodes sharing the same device in such applications. Dual-port support will enable host-side load balancing to this target.

Two controllers can be used in partnership to provide a very high level of fault tolerance. Both controllers share the same back-end SCSI or Fibre Channel devices. When both controllers are in a normal operating state, both are active and load balanced. If one controller fails, the other automatically should invoke a "fail over" allowing the host system to continue to operate normally. When the other controller is replaced or repaired, the controller fails back and maintains active-active load balancing.

A RAID controller that uses Fibre Channel as host interface may support Fibre Channel Switch Fabric (FC-SW) for public loop and point-to-point protocols.

A RAID controller that uses Fibre Channel as the host interface may support Fibre Channel IP protocol (FC-IP) for controller, drive enclosure, and RAID management.

An outboard RAID controller for a subsystem should provide enclosure management services for drive operating status, enclosure temperature, and power voltage. The system should include indications for drive, fan, and power supply operating status, and imminent fault alert indication. The devices should be monitored and managed by the RAID controller.

RAID subsystems should support simple and safe device online plug and unplug for disk device replacement. A hot-swappable drive should have a local indicator that indicates which drives are ready for replacement, facilitating the servicing process and improving reliability by reducing possible errors. For subsystems with multiple drives, an individual replacement indicator should be physically associated with each hot-swappable drive slot. For maximum up time, the RAID subsystem should support hot-swapping capabilities for power supply replacement, power supply redundancy, and hot-swappable cooling fans.

Summary

This is the first in a series of DIG64 technical whitepapers. It is concerned with server mass storage technologies such as SCSI, FC, and RAID, for which interface specifications are published and widely available. The goal of this series of documents is to provide sufficient technical information to designers and integrators of subsystem components for use in IA-64 server systems so that the resulting systems are highly reliable, available, serviceable, and fully performant. It is not the goal of these documents to specify or define operating system policies, application usage characteristics, or the philosophy of the design of IA-64 server subsystem components or devices.

For a more in-depth discussion of each of these technologies, refer to the list of documents in the References section below. In this paper, no consideration has been given to incomplete or in-progress technologies or interface specifications. Future storage technologies will be addressed when those interface specifications are published and made available.

References

ATAPI Removable Media BIOS Specification (ARMD), Version 1.0
http://www.ptltd.com/techs/specs.html

Attachment with Packet Interface Extension (ATA/ATAPI-4) NCTIS 317
http://www.nssn.org

Developer's Interface Guide for IA-64 Servers
http://www.dig64.org

Fibre Channel - Physical and Signaling Interface (FC-PH) ANSI X3.230
http://www.nssn.org/

Fibre Channel - Physical and Signaling Interface (FC-PH) ANSI X3.297
http://www.nssn.org/

Fibre Channel - Physical and Signaling Interface (FC-PH) ANSI X3.303
http://www.nssn.org/

Open Host Controller Interface
http://www.compaq.com/productinfo/development/

PCI Local Bus Specification, Revision 2.1
http://www.pcisig.com/

PCI Local Bus Specification, Revision 2.2
http://www.pcisig.com/

PCI-PCI Bridge Architecture Specification, Revision 1.1
http://www.pcisig.com/

PCI Bus Power Management Interface Specification, Revision 1.1
http://www.pcisig.com/

PCI Hot-Plug Specification, Revision 1.0
http://www.pcisig.com/

Addendum 1.0 to PCI Local Bus Specification,
http://www.pcisig.com/

SCSI1 *Small Computer System Interface-3 (SCSI-3)*
ANSI X3-253
http://www.nssn.org/

Small Computer System Interface-3 (SCSI-3)
ANSI X3-253, Appendix F
http://www.nssn.org/

ANSI NCITS T10 SCSI-3 Multi-Media Command Set-2 (MMC-2)
http://www.symbios.com/x3t10/drafts.htm

Storage Device Class Power Management Reference Specification, Revision 1.0
http://www.microsoft.com/hwdev/onnow.htm#Specs

System Management BIOS Reference Specification
http://developer.intel.com/ial/WfM/design/BIBLIOG.HTM